ADAPTING NOISY SPEECH MODELS – EXTENDED UNCERTAINTY DECODING

Jianhua Lu, Ji Ming, and Roger Woods

Institute of Electronics, Communications and Information Technology
Queen's University Belfast, Queen's Road, Queen's Island, Belfast, BT3 9DT, UK
{jlv01, j.ming, r.woods}@qub.ac.uk

ABSTRACT

Most conventional techniques for noise adaptation assume a clean initial speech model which is adapted to a specific noise condition using adaptation data accumulated from the condition. In this paper, a different problem is considered, i.e. adapting a *noisy* speech model to a specific noise condition. For example, the initial noisy model may be a multi-condition model which is used to provide more accurate transcripts for the adaptation data than could be provided by a clean model, thereby obtaining a more accurate adaptation. We develop the formulation for this new problem by combining and extending maximum likelihood linear regression (MLLR), constrained MLLR (CMLLR) and uncertainty decoding techniques. We also present an implementation which has been tested on the Aurora 4 database, assuming an initial multi-condition model trained using white noise corrupted data. Significant word error rate (WER) reductions are achieved in comparison with other approaches.

Index Terms— noise adaptation, noise compensation, noise robustness, speech recognition, Aurora 4

1. INTRODUCTION

Speaker adaptation approaches, such as MLLR and CMLLR, have been successfully applied to noise adaptation for robust speech recognition [1], [2]. Uncertainty decoding, which extends CMLLR to include the uncertainty of the noise removal, has shown the potential to improve further the noise robustness [3], [4]. Most of these current adaptation techniques are applied to a *clean* speech model, with the aim of transforming this clean model to match a specific noisy environment. The work in this paper addresses a different problem, in which the initial model to be adapted is not a clean speech model, but a *noisy* speech model. For example, this noisy speech model could be a multi-condition model trained using data from a number of expected noise conditions. Multi-condition models are widely used as an alternative to the clean model to offer improved noise robustness, however, they lack sharpness or optimality to a specific noise. Using a multi-condition model as the initial model may benefit from the additional robustness in transcribing the adaptation data.

We extend the uncertainty decoding technique, previously applied to clean initial models, to suit this new problem. As will be shown, the new problem can be formulated as conventional uncertainty decoding plus additional MLLR and transformations to the model's mean vectors and covariance matrices, which account for the difference between the noisy initial model and the target model. To implement the new adaptation method, we propose an algorithm based on subspace distribution clustering [6], which significantly reduces the computational complexity with little effect on the recognition accuracy. The paper concludes with experimental comparisons with other adaptation techniques on the Aurora 4 corpus with a task for large-vocabulary continuous speech recognition in variable noise conditions.

2. CONVENTIONAL APPROACHES TO NOISE ADAPTATION

An initial acoustic model for clean speech, s, is given by (μ_m^s, Σ_m^s) where μ_m^s and Σ_m^s are the mean vector and covariance matrix of the m^{th} Gaussian component, respectively. Given adaptation vectors, y, the conventional MLLR, front-end CMLLR and uncertainty decoding techniques involve adaptation of (μ_m^s, Σ_m^s). These techniques are summarized briefly next, as a starting point for the new problem.

2.1. MLLR

MLLR has been initially developed for speaker adaptation [1] and been particularly effective in noisy speech recognition. In MLLR, the mean vectors of the clean model are adapted using an affine transform which can be expressed as:

$$\mu_{m_r}^y = A_r \mu_{m_r}^s + b_r \tag{1}$$

where r is the index for regression classes, (A_r, b_r) are the transform parameters and $\mu_{m_r}^y$ is the adapted mean vector associated with class r. In equation (1), the initial clean speech model is modified towards the noisy observation, y, by replacing the clean model mean μ_m^s with the adapted mean $\mu_{m_r}^y$.

This work is supported by the UK EPSRC under Grant No. EP/D048605/1.

2.2. Front-end CMLLR

Front-end CMLLR (FE-CMLLR) is applied to remove the noise from the noisy speech vector, y. Assuming that the noise characteristics can be classified into different classes where each class is modeled by a Gaussian e.g. [4], then FE-CMLLR can be expressed as:

$$y_c = A_c y + b_c \tag{2}$$

where c is the index for noise classes, (A_c, b_c) are the transform parameters associated with noise class c and y_c is the corresponding clean speech estimate. The optimal value of c may be identified on the front-end noise model based on the maximum-likelihood principle. In this formulation, the mean vectors and covariance matrices of the initial clean speech model remain unchanged, assuming that the noise in the observation can be removed, as shown in equation (2).

2.3. Uncertainty Decoding

Uncertainty decoding extends CMLLR by additionally taking into account the uncertainty of the noise removal process [3]. It is assumed that both the clean speech and noise are subject to Gaussian mixture models (GMMs). Uncertainty decoding takes the following expression:

$$y_c = A_c y + b_c \tag{3}$$

$$\Sigma_m^{y_c} = \Sigma_m^s + \delta\Sigma_c^e \tag{4}$$

where as in FE-CMLLR, y_c is an estimate of the clean speech vector which by assumption can be expressed as $y_c = s + e_c$ and e_c represents an error associated with noise class c. Assuming that each e_c is a zero-mean random vector with covariance matrix $\delta\Sigma_c^e$, we then get equation (4) where $\Sigma_m^{y_c}$ is the adapted covariance matrix taking into account the uncertainty of the clean speech estimate caused by e_c. The m^{th} mixture component can thus be written as:

$$\begin{aligned} p(y|m) &\approx p(y|m,c) \\ &= |A_c|\, N(y_c; \mu_m^s, \Sigma_m^{y_c}) \end{aligned} \tag{5}$$

This formulation includes filtering the noisy observation and replacing the clean model covariance Σ_m^s with the adapted $\Sigma_m^{y_c}$, but it leaves the clean model mean μ_m^s unchanged, since it is assumed that e_c is a zero-mean random vector.

3. A NEW PROBLEM AND THE FORMULATION

The above methods focus on adapting a clean speech model (μ_m^s, Σ_m^s) to noisy observations subject to a specific noise condition. In this paper, we consider a different problem – adapting a *noisy* speech model (μ_m^x, Σ_m^x) to y, where x represents a noise condition different from y or not as specific as the noise condition in y. The model (μ_m^x, Σ_m^x), for example, could be a multi-condition model trained using data from a number of different noise conditions. A multi-condition model can be used as an alternative initial model to the clean model for adaptation. Although lacking the optimality to a specific noise, it may provide more robust transcripts for the adaptation data than can be provided by the clean model, and thus lead to a more accurate adaptation. Additionally, (μ_m^x, Σ_m^x) could be a model trained earlier for a specific type of noise, and now needs to be adapted to a new type of noise. In this case, the previous methods for adapting clean speech models may not be directly applicable. Here we propose a solution, which extends the uncertainty decoding and MLLR by taking into account the difference between the model to be adapted (μ_m^x, Σ_m^x) and the clean model (μ_m^s, Σ_m^s). The new formulation is written as:

$$y_c = A_c y + b_c \tag{6}$$

$$\Sigma_{m_r}^{y_c} = \Sigma_{m_r}^x + \delta\Sigma_c^e + \delta\Sigma_r^{s-x} \tag{7}$$

$$\mu_{m_r}^{y_c} = C_r \mu_{m_r}^x + d_r. \tag{8}$$

where as with equations (3) and (4), c addresses the front-end noise class, y_c is an estimate of the clean speech, and $\delta\Sigma_c^e$ accounts for the uncertainty of the estimation. In equation (7), an adjustment $\delta\Sigma_r^{s-x}$ is included to reflect the covariance difference between the clean speech s and the training speech x, in terms of $\Sigma_{m_r}^x + \delta\Sigma_r^{s-x} \simeq \Sigma_{m_r}^s$. Equation (8) adapts the mean vector towards the clean speech estimate y_c from the noisy mean vector $\mu_{m_r}^x$, using an MLLR formulation with (C_r, d_r) being the parameters of regression class r. The m_r^{th} mixture component for noisy observation, y, is expressed as:

$$\begin{aligned} p(y|m_r) &\approx p(y|m_r, c) \\ &= |A_c|\, N(y_c; \mu_{m_r}^{y_c}, \Sigma_{m_r}^{y_c}) \end{aligned} \tag{9}$$

As shown in equations (6)–(8), the model includes front-end and back-end transform parameters, $\mathcal{F} = (A_c, b_c, \delta\Sigma_c^e)$ and $\mathcal{B} = (C_r, d_r, \delta\Sigma_r^{s-x})$ respectively. $\mathcal{F}$ can be estimated using the methods as used in conventional front-end based uncertainty decoding [4]. Given $\mathcal{F}$, an estimate of $\mathcal{B}$ can be obtained by minimizing the auxiliary function:

$$\begin{aligned} Q(\mathcal{B}', \mathcal{B}) = \sum_t \sum_r \sum_{m_r} \gamma_t(m_r) \Big[&\log|\Sigma_{m_r}^{y_c}| + \\ &(y_c(t) - \mu_{m_r}^{y_c})^{\mathsf{T}} (\Sigma_{m_r}^{y_c})^{-1} (y_c(t) - \mu_{m_r}^{y_c}) \Big] \end{aligned} \tag{10}$$

where $y_c(t)$ represents the estimated frame vector at time t and $\gamma_t(m_r)$ is the occupation probability of mixture component m_r by $y_c(t)$. Minimizing the auxiliary function w.r.t (C_r, d_r) is the same as MLLR, however, solving equation (10) w.r.t $\delta\Sigma_r^{s-x}$ may be a difficult task given that a numerical method would normally be needed. The complexity of the numerical method increases with the number of Gaussians in the model. For over 10,000 Gaussians as normally seen in large-vocabulary systems, this is impractical. In the following, we present an approach to simplify this

computation by directly reducing the number of Gaussians in the model.

4. IMPLEMENTATION

In our implementation, we compress all the Gaussians to a set of codewords and then minimize the codeword-based model, equation (10) w.r.t $\delta\Sigma_r^{s-x}$ using a Newton interval method.

4.1. Model Compression

Assume that the feature vector, x, can be divided into K mutually independent streams, i.e. $x = [\ x_1^{\mathsf{T}} \quad x_2^{\mathsf{T}} \quad ... \quad x_K^{\mathsf{T}}\ ]^{\mathsf{T}}$. If x is modeled by a Gaussian mixture model with diagonal covariance matrices, then the m^{th} Gaussian component can be written as:

$$p(x|m) = \prod_{k=1}^{K} N(x_k; \mu_{mk}^x, \Sigma_{mk}^x) \tag{11}$$

where $(\mu_{mk}^x, \Sigma_{mk}^x)$ is the k^{th} stream Gaussian in the m^{th} mixture component. If some of the stream Gaussians are similar to each other (measured by the Bhattacharyya distance [6], for example), they can be merged into one Gaussian codeword $(\mu_{ik}^x, \Sigma_{ik}^x)$, addressed by (i,k), where

$$(i,k) = f(m,k) \quad 1 \le m \le M; 1 \le i \le I \tag{12}$$

where f is a mapping function converting (m,k) to (i,k), M is the total number of mixtures in the original model, and I is the number of the codewords for each stream (assuming an equal number of codewords per stream). Assume that equation (11) can be approximated on the codewords, the following codeword-based model is obtained:

$$p(x|m) \approx \prod_{k=1}^{K} N(x_k; \mu_{ik}^x, \Sigma_{ik}^x) \tag{13}$$

which shares the same principle with the subspace distribution clustering (SDC) HMM [6]. Our early work indicates that for a typical 5k-word recognition task, the number of stream Gaussians can be reduced from $M = 51,856$ to $I = 512$ per stream, with little degradation in recognition accuracy [5].

4.2. Regression on Codeword-based Model

For simplicity, assume a global font-end noise class is used in equation (6), with parameters (A,b), so that the feature transformation can be written as $y^* = Ay + b$. Let $(\mu_{i_rk}^x, \Sigma_{i_rk}^x)$ denote a codeword of the k^{th} stream belonging to regression class r and allow regressions be applied independently within each stream instead of to the whole frame vector as in equation (1). The regression formulae in equations (7) and (8) can be realized, within each individual stream, as follows:

$$\mu_{i_rk}^{y^*} = C_{rk}\mu_{i_rk}^x + d_{rk} \tag{14}$$

$$\Sigma_{i_rk}^{y^*} = \Sigma_{i_rk}^x + \delta\Sigma_{rk}^x \tag{15}$$

where (C_{rk}, d_{rk}) are the regression parameters for the k^{th} stream mean vector, and $\delta\Sigma_{rk}^x$ is an estimate of the sum of the covariance biases $\delta\Sigma_{ck}^e$ and $\delta\Sigma_{rk}^{s-x}$ for the k^{th} stream. Assuming that (A,b) are included as part of the model parameters to be estimated along with the other model parameters, a new model transform parameter is defined as $\mathcal{T} = (A, b, C_{rk}, d_{rk}, \delta\Sigma_{rk}^x)$, and the occupation probability of codeword (i,k) as

$$\gamma_t(i,k) = \sum_{m:f(m,k)=(i,k)} \gamma_t(m,k) \tag{16}$$

Substituting equation (16) into equation (10), approximating $(\mu_{m_r}^x, \Sigma_{m_r}^x)$ with the corresponding codewords, and taking into account the newly added model-based front-end parameters (A,b), equation (10) is rewritten as:

$$\begin{aligned} &Q(\mathcal{T}', \mathcal{T}) = \\ &T\log|A| + \sum_t \sum_r \sum_{(i_r,k)} \gamma_t(i_r,k)\Big[\log|\Sigma_{i_rk}^x + \delta\Sigma_{rk}^x| \\ &+ (y_k^*(t) - \mu_{i_rk}^{y^*})^{\mathsf{T}}(\Sigma_{i_rk}^x + \delta\Sigma_{rk}^x)^{-1}(y_k^*(t) - \mu_{i_rk}^{y^*})\Big] \end{aligned} \tag{17}$$

where T is the number of the frames used for adaptation. Estimating (A,b), (C_{rk}, d_{rk}) and $\delta\Sigma_{rk}^x$ can be alternated towards a convergence. The standard CMLLR and MLLR techniques can be used, respectively, to estimate (A,b) and (C_{rk}, d_{rk}). Given (A, b, C_{rk}, d_{rk}), we can use a numerical method to find an optimal $\delta\Sigma_{rk}^x$. In this paper, we first find an interval $(\delta\Sigma_1, \delta\Sigma_2]$ for $\delta\Sigma_{rk}^x$ that satisfies:

$$\frac{\partial Q(\mathcal{T}', \mathcal{T})}{\partial \delta\Sigma_{rk}^x}\Big|_{\delta\Sigma_{rk}^x = \delta\Sigma_1} \cdot \frac{\partial Q(\mathcal{T}', \mathcal{T})}{\partial \delta\Sigma_{rk}^x}\Big|_{\delta\Sigma_{rk}^x = \delta\Sigma_2} < 0 \tag{18}$$

The Newton interval method is then used to search within $(\delta\Sigma_1, \delta\Sigma_2]$ the root of $\partial Q/\partial\delta\Sigma_{rk}^x = 0$. This method has a complexity of $\mathcal{O}(I)$, and has been found to be effective in our experiments on the codeword-based model in terms both of low computational cost and of recognition accuracy.

5. EXPERIMENTAL RESULTS

The Aurora 4 database [8], for large-vocabulary continuous speech recognition in noise conditions, was used for the experiments. The test data is divided into seven sets (labelled 1–7), each containing with a different background: clean, car, babble, restaurant, street, airport and train and sampled at 16 kHz (wv1). For training, 7138 clean utterances are provided. To evaluate the new adaptation method, we created a multi-condition model as an initial model, trained using the clean

Method	Test set							Average
	1	2	3	4	5	6	7	
SPLICE UD	14.5	17.5	23.5	32.5	24.0	23.1	26.1	23.0
Mask UD [7]	10.5	14.0	20.0	22.0	24.9	17.5	25.7	19.2
SUBREST [5]	10.1	13.4	20.0	21.4	20.9	17.8	23.8	18.2
New	10.0	11.8	16.6	20.1	20.0	16.9	22.3	16.9

Table 1. WER (%) on Aurora 4, comparing the proposed new adaptation method with uncertainty decoding (UD) and SUBREST. The results for the Mask UD and SUBREST were quoted from [7] and [5], respectively.

training data plus artificial noisy data created by adding white noise at five different signal-to-noise rates (SNRs): 20dB, 15dB, 10dB, 5dB and 0dB. The model consisted of 3241 states with 16 mixtures per state. This multi-condition model used subband-based, 30-stream frame vectors followed by missing-feature decoding to offer an early-stage robustness to transcribe the adaptation data [9], [5]. Using the scheme described in Section 4.1, the whole model, with 3241 states $\times$ 16 mixtures $\times$ 30 streams, was compressed to a codebook consisting of 512 codewords for each of the 30 streams.

The proposed adaptation is conducted in a unsupervised and incremental fashion. For comparison, two uncertainty decoding systems were used. The first is based a SPLICE front-end [3], which is applied to the same multi-condition model as described above, trained using the artificial noisy training data; the second is based on a binary-mask front-end applied to a clean initial model [7]. A third system, namely SUBREST (subband re-estimation) reported in an early paper [5], was included in the comparison. The new method described in this paper extends SUBREST by using the regression (i.e. (7), (8)) to replace the model parameter re-estimation. Compared to the UD, the proposed method has twice the number of parameters in the UD.

Table 1 shows the word error rate comparisons. The new system outperformed the others with 26.5/12.0/7.1% WER reductions to the SPLICE UD, Mask UD and SUBREST, respectively. As the SPLICE UD was applied to the same multi-condition model as with the proposed, there could be two reasons for the performance variation: the separation of the front-end processing from the back-end decoding, and the omission of the mean vector adaptation, i.e. equation (8) in the SPLICE UD. The new system performed better than the Mask UD, possibly due to the use of a multi-condition initial model to transcribe the adaptation data. The new system also outperformed the SUBREST system due to the more effective use of the adaptation data in the new regression formulation.

6. CONCLUSION

A new noise adaptation technique has been presented for adapting noisy speech models to a new noise environment. The proposed method includes transforms for both the noisy observations and the covariance matrices as in normal uncertainty decoding, and a further transform for the mean vectors. Thus, the new method includes two sets of transforms, one is estimated from the front-end, the other from the back-end. Experiments on the Aurora 4 database have shown significant WER reduction in comparison to other adaptation approaches. The mathematical framework for adapting noisy acoustic model will be investigated in the future work.

7. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, pp. 171–185, 1995.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1997.

[3] J. Droppo, A. Acero, and Li Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," *ICASSP'2002*, pp. 57–60.

[4] H. Liao and M.J.F. Gales, "Joint uncertainty decoding for noise robust speech recognition," *Interspeech'2005*.

[5] J. Lu, J. Ming, and R. Woods, "Replacing uncertainty decoding with subband re-estimation for large vocabulary speech recognition in noise," *Interspeech'2009*.

[6] E. Bocchieri and B.K.-W. Mak, "Subspace distribution clustering hidden Markov model," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 264–275, 2001.

[7] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, pp. 2130–2140, 2007.

[8] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," AU/417/02, 2002.

[9] J. Ming, "Noise compensation for speech recognition with arbitrary additive noise," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, pp. 833–844, 2006.